Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management

Liuzixuan Lin
University of Chicago
Chicago, IL, USA
lzixuan@uchicago.edu

Rajini Wijayawardana
University of Chicago
Chicago, IL, USA
rajini@uchicago.edu

Varsha Rao
University of Chicago
Chicago, IL, USA
varsharao@uchicago.edu

Hai Nguyen
University of Chicago
Chicago, IL, USA
ndhai@cs.uchicago.edu

Wedan Emmanuel Gnibga
Univ. Rennes, Inria, CNRS, IRISA
Rennes, France
wedan-emmanuel.gnibga@irisa.fr

Andrew A. Chien
University of Chicago & Argonne National Laboratory
Chicago, IL, USA
aachien@uchicago.edu

ABSTRACT

The unprecedented rapid growth of computing demand for AI is projected to increase global annual datacenter (DC) growth from 7.2% to 11.3%. We project the 5-year AI DC demand for several power grids and assess whether they will allow desired AI growth (resource adequacy). If not, several "desperate measures"—grid policies that enable more load growth and maintain grid reliability by sacrificing new DC reliability are considered.

We find that two DC hotspots—EirGrid (Ireland) and Dominion (US)—will have difficulty accommodating new DCs needed by the AI growth. In EirGrid, relaxing new DC reliability guarantees increases the power available to 1.6x–4.1x while maintaining 99.6% actual power availability for the new DCs, sufficient for the 5-year AI demand. In Dominion, relaxing reliability guarantees increases available DC capacity similarly (1.5x–4.6x) but not enough for the 5-year AI demand. New DCs only receive 89% power availability. Study of other US power grids—SPP, CAISO, ERCOT—shows that sufficient capacity exists for the projected AI load growth.

Our results suggest the need to rethink adequacy assessment and also grid planning and management. New research opportunities include coordinated planning, reliability models that incorporate load flexibility, and adaptive load abstractions.

CCS CONCEPTS

• **Applied computing** → **Data centers**; • **Hardware** → **Power and energy**; • **Computing methodologies** → *Artificial intelligence.*

KEYWORDS

Power grid, Data centers, Resource adequacy, AI

ACM Reference Format:
Liuzixuan Lin, Rajini Wijayawardana, Varsha Rao, Hai Nguyen, Wedan Emmanuel Gnibga, and Andrew A. Chien. 2024. Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management. In *Proceedings of The 15th ACM International Conference on Future and Sustainable Energy Systems (e-Energy '24).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3632775.3661959

1 INTRODUCTION

The phenomenon of ChatGPT has led to an explosion of interest and use of generative AI chatbots and intelligent assistant services based on large language models (LLMs). LLMs have been incorporated into many widely used computing services such as Slack, Microsoft Office and stimulated new tools like Copilot for programming, etc. Training and operating these large models is costly [7, 52], and the "AI gold rush" has triggered massive investment in GPU hardware and datacenters [47, 48]. For example, the three largest hyperscalers (Amazon, Microsoft, and Google) reported large increases in datacenter capex investment from $78B (2022) to $120B (2023), a $42B or 54% increase [16]. There is increasing awareness of exponential datacenter growth due to AI and resulting energy crisis [2, 4, 11, 15, 30, 44, 49, 50]: the existing and planned power grid infrastructure cannot support it, leading to climate-unfriendly measures (delaying shutdown of fossil-based generation). However, this awareness focuses on total power load growth, not specific grid challenges.

Datacenters' rapid load growth is outracing both annual power grid planning cycles and investment/construction cycles (5–20 years) [12, 17, 21, 33]. A number of power grids have halted or slowed new datacenter connections to ensure grid reliability [28, 29]. For power grids, the situation is compounded by additional stresses from extreme climate events and increased generation volatility from increasing dependence on renewables [24, 31, 55]. Therefore, to better understand the challenges and explore new research opportunities, it is necessary to assess the demand growth against *grid resource adequacy*, the method by which power grids decide how much load can be safely connected to the grid.

In this paper, we study a diverse set of power grids—EirGrid (Ireland), Dominion, CAISO, ERCOT, and SPP (US), assessing whether they have enough reliable capacity to meet the 5-year AI/Cloud datacenter power growth. We examine how relaxing new datacenter power reliability guarantees could increase the available grid power for new datacenters. Specific insights include:

e-Energy '24, June 4–7, 2024, Singapore, Singapore
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0480-2/24/06
https://doi.org/10.1145/3632775.3661959

- In EirGrid, we project that AI will increase the DC compound annual growth rate (CAGR) to 16.7%. Its resource plan cannot accommodate the projected load in 2024. Relaxing power reliability for new datacenters increases the power available for new datacenters to 1.6x–4.1x so the AI demand can be met for the next 5 years. New datacenters can expect 99.6% power availability despite lower guarantees.
- Dominion grid, which faces a similar rapid DC growth challenge (23.6% CAGR), cannot meet the growing demand even with reduced new datacenter power reliability. The new datacenters will experience poor power availability (< 90%).
- Study of several other power grids (CAISO, ERCOT, and SPP) finds that they have excess capacity to accommodate AI load growth within reliable grid operation.
- We identify new research opportunities to rethink adequacy assessment, reinvent grid planning and management as cooperative, design new reliability models that incorporate load flexibility, and adaptive load abstractions.

2 PROBLEM

Consider EirGrid, a hotspot for datacenters with 600 MW datacenter load (14% of grid load) in 2021 [21]. First, we project Cloud growth and Cloud+AI growth using the methodology in Section 3.1.2. For Cloud, EirGrid's plan expected a 10.6% CAGR (compound annual growth rate) DC growth (brown line), and AI projections increase the growth rate to 16.7% (red line) as shown in Figure 1.

Using the grid resource adequacy assessment methodology in Section 3.1, we compute the reliable grid limit for 2023-2028 (see [21]). The results show that for Cloud, demand will exceed the reliable grid limit by 2025, and with the more rapid growth of Cloud+AI, demand will exceed the limit even earlier, by 2024.

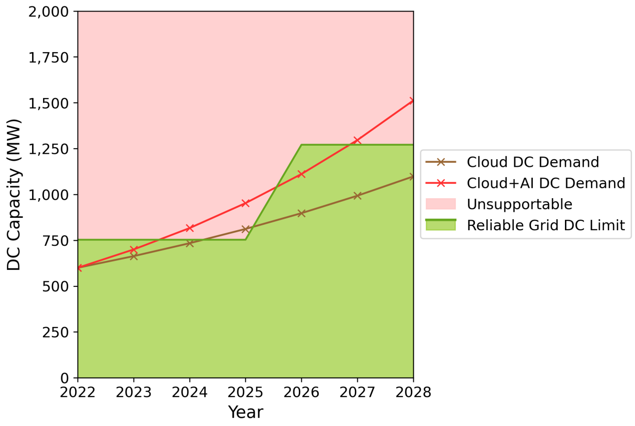


Figure 1: EirGrid DC load is growing rapidly, and AI accelerates it. By 2025, Cloud DC growth cannot be supported, and by 2024, Cloud+AI DC growth cannot be supported.

The shape of the reliable DC limit is determined by changes in generators available on the EirGrid over time (Figure 2). From 2023 to 2026, EirGrid plans to replace 1,500 MW of conventional generation with renewables. This maintains total power available, but increases volatility. Consequently, the reliable load limit decreases, and cannot support the increased datacenter demand.

In 2026, the addition of 800 MW of gas generation lifts the reliable grid limit. A further planned addition of 1,000 MW wind generation in 2028 increases the reliable load limit only marginally. Thus capacity available for new DC load is reduced by substitution of renewables (for conventional generation), and only partially met by new gas generation. To conclude, EirGrid will not be able to support growing DC load by as early as 2024, and certainly by 2025.

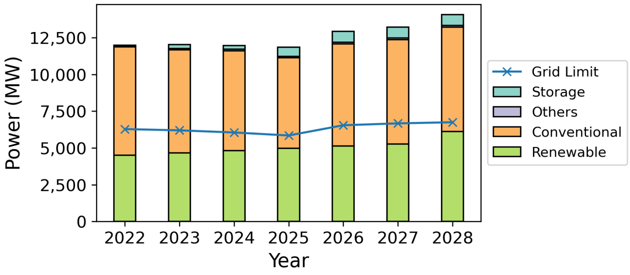


Figure 2: EirGrid's Reliable Load Limit (blue line) and Resources, 2022–2028. The load limit does not increase despite a 30% increase in renewable generation capacity.

3 STUDY: EIRGRID

We describe the resource adequacy (RA) assessment methodology for EirGrid. Then, we consider two different approaches to increase the supportable datacenter load without endangering power grid reliability. These methods require no physical changes, and thus can be applied immediately. Using these methods, we show how they could accommodate growing Cloud+AI datacenter load, and achieve high datacenter availability.

3.1 Methodology

3.1.1 Resource Adequacy Assessment. We use a grid resource adequacy assessment framework based on the published methodologies of a variety of power grids [8, 17, 21]. With grid current and planned (2022–2028) generation resources (conventional and renewables, energy storage, and others such as imports), load and renewable generation traces are used to assess whether grid load can be met for each time interval in a year. Simulating over a year captures the day-to-day variations in load and renewable generation, and their relationship. To model future grid scenarios, 2022's non-DC load and renewable generation traces are scaled proportionally based on expected load growth and planned generation capacity, respectively. The result of resource adequacy assessment is the **loss of load expectation (LOLE)** metric, whose definition and standard vary by grid. For EirGrid, LOLE is the expected number of hours in a year when the available grid capacity cannot meet load, and the target is ≤ 8 hours per year (99.9% reliability). Formally, for year y:

$$LOLE_y = (\sum_{t=1}^{T} I_{y,t}) * intLen/60 \tag{1}$$

where T is the number of data time intervals in a year and $intLen$ is the length of time interval in minutes (15 minutes for EirGrid). I_t indicates whether capacity shortage happens at time t:

$$I_{y,t} = \begin{cases} 1, & load_{y,t} > renewGen_{y,t} + convCap_y + stoCap_y + other_y \\ 0, & otherwise \end{cases} \tag{2}$$

$$load_{y,t} = load_{DC,y,t} + load_{NonDC,y,t} \tag{3}$$

where $renewGen_{y,t}$ is the actual renewable generation, and $convCap_y$, $stoCap_y$, and $other_y$ are capacities of conventional generation, energy storage, and other resources de-rated by availability factors (details in Appendix A). The grid datacenter limit is defined as the maximum datacenter load that produces LOLE $\leq$ 8 hours.

The major uncertainties that this method doesn't model are year-to-year weather changes and load profile reshaping with load flexibility. These are both opportunities for future research (see Section 6).

3.1.2 Datacenter Load Growth Model. Projected Cloud datacenter power demand is based on EirGrid's forecast of Cloud growth of 10.6% CAGR [21]. Cloud+AI demand adds the incremental AI demand estimated by extrapolating NVIDIA's latest datacenter revenue guidance released in August 2023 [40]. This AI demand estimation produces 1,400 MW additional global datacenter capacity (4.1% of current capacity) next year (details in Appendix B). Because datacenter growth is not uniform geographically, we make a location-based adjustment that assumes more new capacity will be deployed in higher-value regions (e.g. EirGrid, Dominion) based on growth rates for Cloud alone. For region r,

$$CAGR_{cloud+AI,r} = CAGR_{cloud,r} + \frac{CAGR_{cloud,r}}{CAGR_{cloud,avg}} \cdot \Delta CAGR_{AI,avg}$$

with $CAGR_{cloud,avg}$ (global average) of 7.2%, the datacenter demand growth including AI comes to 16.7% CAGR for EirGrid.

3.1.3 New Schemes to Accommodate Datacenter Load. To avoid threatening grid reliability, researchers and, more recently, commercial players have proposed unreliable power service for datacenters [9, 41, 56], in effect, reducing the quality-of-service (QoS) for datacenter power from reliable to partially reliable. So here, we consider datacenter grid connection with reduced reliability, and examine what the consequences are for the grid and the datacenters—outages. **Datacenter outage** is defined as time periods when datacenter load is shed by the grid, and we report the outage rate.

We study three different levels of QoS for new datacenters:

- **Reliable.** The usual assumption in grid resource adequacy assessment [21]. Datacenters get reliable, continuous power supply up to maximum capacity.
- **80% Reliable.** Datacenters get reliable, continuous power supply up to 80% of the capacity. The 20% is only available when it doesn't tax the grid (see [9]).
- **0% Reliable.** Datacenters get reliable, continuous power supply for 0% of the capacity. That is, no guaranteed power. The 100% is only available when it doesn't tax the grid. The outage rate is limited to 1% (88 hours in a year). An example is ERCOT's Large Flexible Loads program [41] or the ideas espoused in Zero-carbon Cloud [56].

The benefit of reducing QoS for new datacenters is more DC load can be safely attached to the grid. If the grid is overloaded, the DC load can be reduced, protecting grid reliability while avoiding grid service violations as these reductions do not count as LOLE.

3.2 Results

We compute resource adequacy limits under the reduced DC QoS schemes, and plot the results in Figure 3. Note that Figure 3 shows annual capacity limits, differing from Figure 1 which shows stable capacity available into the future (constrained by future minimum).

Relaxing new DC QoS to 80% reliable enables EirGrid to accommodate 25% greater datacenter capacity, a significant improvement, but still insufficient to support the increasing AI demand. Further reducing new DC QoS to 0% reliable increases the available new datacenter capacity to 1.6x–4.1x, enabling the EirGrid to support ALL of the projected Cloud+AI demand through 2028. Note that the demand increase from 2023 is more than 100% (2x).

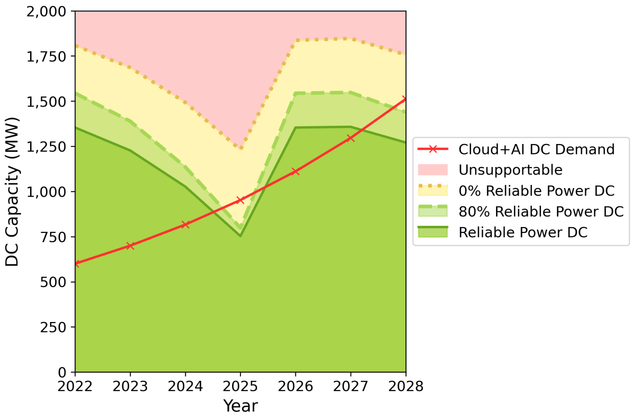


Figure 3: EirGrid Grid Datacenter Capacity Limits with Varied New Datacenter Power QoS vs. Demand.

How often does the grid fail to fully power the reduced QoS datacenters? We computed the EirGrid daily datacenter outage rate for 2028 with 0% reliable new datacenter QoS (Figure 4). Each bar reports the daily fraction of time when there is datacenter load shedding. Shedding occurs primarily in winter (beginning and end of year) and in total is 34.5 hours in a year, and typically a small magnitude. So despite no guarantee, the new datacenters actually experience power supply QoS of 99.6%, nearly the grid LOLE goal of 99.9%. We expect that with in-advance grid alerts [22], the negative impacts on datacenter operation should be minimal. This suggests research on dynamic adequacy (dynamic models for reliability) is an interesting opportunity.

4 OTHER GRIDS

We continue assessing whether the DC load growth can be met in the other four power grids (Dominion, CAISO, ERCOT, SPP). These grids vary in size, load, and generation mix, and have well-documented resource plan and datacenter capacity (Appendix A, B).

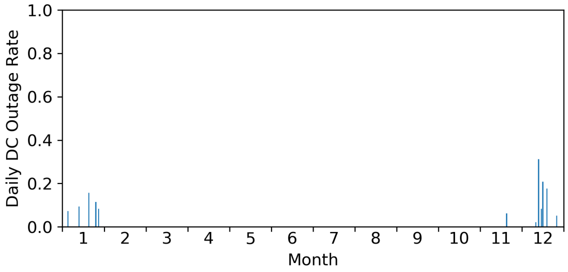


Figure 4: Datacenter Outage Rates (daily fraction), 2028 EirGrid: using 0% reliable DCs to meet 900 MW AI demand. Outages are rare (0.4%, 34.5 hours) and seasonal (winter).

With proximity to the US government and transatlantic networking, Northern Virginia is home to the world's highest concentration of data centers [5, 20], whose power demand is mainly served by Dominion Energy. This datacenter concentration is expected to continue, and its growth has already exceeded grid operator's expectation several times, as reflected in Dominion's repeated revisions on load forecast [17]. In 2022, Dominion forecast 15% CAGR datacenter load growth [12] before 2028, and we project (Section 3.1.2) with AI growth this DC growth will rise to 23.6% CAGR.

Figure 5 compares the grid datacenter capacity limits and the demand growth. The projection shows that the existing 2,767 MW datacenter capacity [17] will grow to 9,800 MW in 2028. However, the reliable datacenter limit under Dominion's planned resources [17] is only 55% of the demand. Relaxing new datacenter QoS to 80% reliable fails to meet DC demand. Going further to 0% reliable increases the limit but still only 70% of the demand. With 100% demand, datacenters added under 0% guarantee (Figure 6) will experience frequent outages (1,010 hours in a year, 11.5% time) unacceptable to datacenter operators. In summary, for Dominion, relaxing new DC QoS increases capacity, but cannot meet the growing AI DC load. Actually, it already has shortfalls [28], and new research on cooperative capacity planning (and investment) is needed (see Section 6).

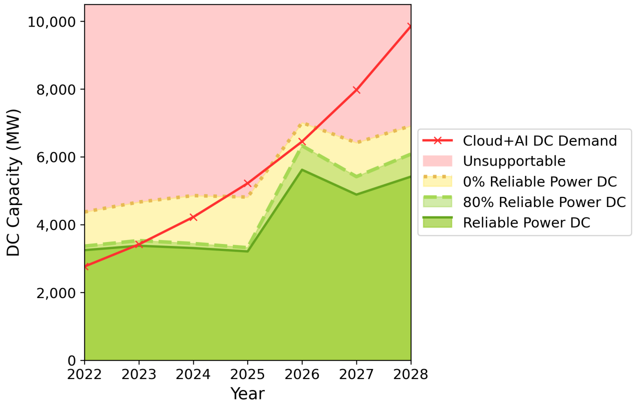


Figure 5: Dominion Grid Datacenter Capacity Limits with Varied New Datacenter Power QoS vs. Demand.

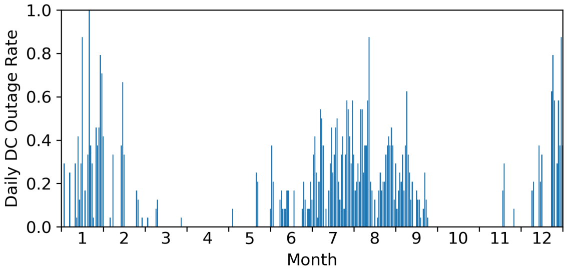


Figure 6: Outage Rates (daily fraction), 2028 Dominion Grid: using 0% reliable datacenters to meet 7,100 MW AI-demand. Winter and Summer see 1,010 hours' outage in 145 days.

For the other three grids, excess datacenter capacity (available reliable capacity minus demand) will exist for the next 5 years (Figure 7): new datacenters can get 100% reliable power supply in these grids. The reasons for excess datacenter capacity vary by grid.

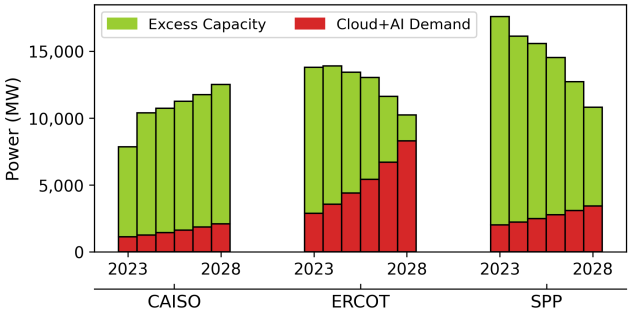


Figure 7: Excess Datacenter Capacity in CAISO, ERCOT, and SPP for 2023–2028, despite increasing Cloud and AI demand.

CAISO (California, US) has the most aggressive decarbonizaton plan, targeting 60% RPS by 2030 and 90% by 2035 [14]. It plans to retire 6,000 MW conventional and add 12,000 MW solar and 9,000 MW wind by 2028 [8]. CAISO's strategy can bring 73% [26] of the total 15,296 MW energy storage online for resource adequacy during peak hours.

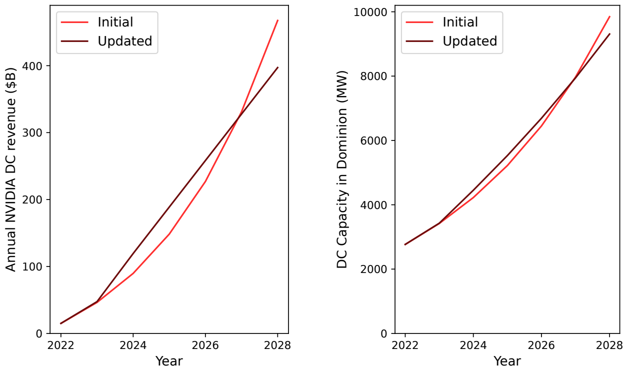
ERCOT (Texas, US) plans to add 30,000 MW solar and 11,000 MW energy storage by 2028 [10, 43] to support energy demand growth of 23% (industry growth) [42]. Grid resources are adequate for datacenter growth if we use CAISO computation for energy storage contribution, but under ERCOT's pessimistic assumption (0 contribution during peak hours), it might not meet AI demand. Further, as an isolated grid, it faces risk during extreme events [32].

SPP (Southwest, US) has a large reserve margin (20%) in generation resources as of 2023 and will maintain sufficient conventional generation balancing additions and retirements by 2028 [45]. This conservative plan meets projected demand.

5 REVISITING AI POWER GROWTH AND GRID READINESS

In the six months since we began this study in September 2023, concern about AI's increasing power demand and the threat it represents to grid stability and decarbonization has become a headline concern [44]. In this section, we update our demand growth projection and grid readiness as of March 2024.

Continued growth in AI infrastructure [18, 51], and growth of NVIDIA's datacenter revenue indicate that AI power consumption growth is accelerating. Considering NVIDIA's latest actual and projected datacenter revenue (Appendix B.1), the four most recent quarters produce a 12-month total revenue of $66.6B. The trend suggests linear datacenter revenue growth. Thus, we propose a linear DC growth model, less aggressive than our initial model (exponential, Section 3.1.2) and evaluate its effects. Based on NVIDIA's average $4.34B quarterly datacenter revenue increase, we project through 2028. The resulting annual datacenter revenue and corresponding power load are shown in Figure 8. Compared with our previous model, the new model reflects a slightly higher load in the near-term and a lower load in the long-term. We consider this projection more realistic.



(a) NVIDIA's Annual DC Revenue **(b) Dominion Grid DC Demand**

Figure 8: Comparison between the Updated (linear) and Initial (exponential) Datacenter Projections.

We compare the updated Dominion load growth projection to the capacity limits in Figure 9. Dominion has not released its 2024 Integrated Resource Plan, so we see higher demand against the same capacity limits. Our conclusions are essentially the same with updated demand: shortage begins in 2024, and relaxing new datacenter QoS to 0% reliable can still only meet 74% of the demand in 2028. For the other four grids we studied, EirGrid added more renewable generation and energy storage resources in accordance with Irish government's updated goal of 80% (was 70%) renewable electricity by 2030 [23]. These resources can increase the reliable datacenter capacity limit in EirGrid to meet the demand. We have not found any changes in the other three US grid (CAISO, ERCOT, SPP) plans, perhaps because they are expected to have excess datacenter capacity in our assessment. However, rising awareness of energy demand growth challenges [27] suggests the situation may change in the future as resource adequacy concerns are spreading to more grids [13].

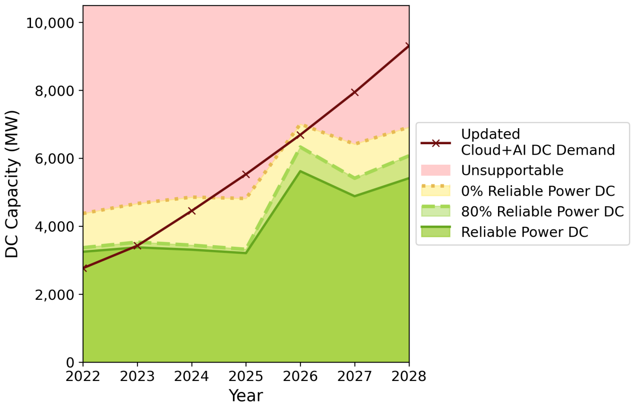


Figure 9: Dominion Grid Datacenter Capacity Limits vs. Updated Demand.

6 RESEARCH OPPORTUNITIES

The insight that the grid infrastructure cannot meet rapid Cloud+AI growth opens many research opportunities—rapid change is essential to solve this and other renewable grid problems. We describe a few exciting new research opportunities below.

Coordinated Planning and Co-investment. Traditional extrapolation [17, 21] is insufficient to track exponential growth (e.g. the AI boom), see [12, 13]. Now that datacenter loads are large, we need new information sharing and co-investment (de-risking) by wealthy computing companies to ensure sufficient power infrastructure for 3, 7, even 10-year windows.

New Models for Reliability. Traditional statistical models for reliability were appropriate for loosely coupled, uncontrollable loads. With large fluctuations in renewables and advanced load control beyond traditional demand-response [41], how can we assess adequate supply and reliability? Quantity of generation, ramping, etc. are not enough for tightly coupled, dynamically controlled systems.

New Abstractions for Load and Adaptive Load Management. Recent planning reports for renewable-dominated grids exploit terms such as "Shape, Shift, Shed, and Shimmy" [19], to describe load properties. Should datacenters describe their load in these terms—to shape computing resource management and QoS for services? [37, 46, 54, 56, 57] Can this enable cooperative datacenter and grid planning? [34, 36]

ACKNOWLEDGMENTS

We thank the anonymous reviewers for the insightful reviews. This work is supported in part by NSF Grant CNS-1901466, and the VMware University Research Fund. We also thank the Large-scale Sustainable Systems Group members for their support of this work!

REFERENCES

[1] International Energy Agency. 2023. Data centers & networks - IEA. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

[2] Aventine. 2024. What Can We Do About AI's Insatiable Thirst for Energy? https://www.aventine.org/energy-consumption-ai-lab-grown-meat

[3] Baxtel. 2023. United States Data Center Market. https://baxtel.com/data-center/united-states

[4] Justine Calma. 2024. AI and crypto mining are driving up data centers' energy use. *The Verge* (Jan 2024). https://www.theverge.com/2024/1/24/24049047/data-center-ai-crypto-bitcoin-mining-electricity-report-iea

[5] CBRE. 2023. Global Data Center Trends 2023. https://www.cbre.com/insights/reports/global-data-center-trends-2023.

[6] Andrew A Chien. 2023. GenAI: Giga ,TeraWatt-Hours,andGigaTonsofCO2. *Commun. ACM* 66, 8 (2023), 5–5.

[7] Andrew A Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems* (Boston, MA, USA) *(HotCarbon '23)*. Association for Computing Machinery, New York, NY, USA, Article 11, 7 pages. https://doi.org/10.1145/3604930.3605705

[8] California Public Utilities Commission. 2023. Proposed Electricity Resource Portfolios for the 2023-2024 Transmission Planning Process. https://www.cpuc.ca.gov/-/media/cpuc-website/divisions/energy-division/documents/integrated-resource-plan-and-long-term-procurement-plan-irp-ltpp/2022-irp-cycle-events-and-materials/2023-2024-tpp-portfolios-and-modeling-assumptions/23-24tpp_portfolios_workshopslides.pdf

[9] Ireland Commission for Regulation of Utilities. 2021. CRU proposed Direction to the System Operators related to Data Centre grid connection. https://cruie-live-96ca64acab2247eca8a850a7e54b-5b34f62.divio-media.com/documents/CRU21060-CRU-consultation-on-Data-Centre-measures.pdf.

[10] Astrapé Consulting. 2023. 2022 Zonal Reliability Study. https://www.ercot.com/files/docs/2023/01/10/ERCOT_Zonal_Reliability_Study_Report_1-9-2023.pdf

[11] Alex de Vries. 2023. The growing energy footprint of artificial intelligence. *Joule* 7, 10 (2023), 2191–2194.

[12] PJM Resource Adequacy Planning Department. 2023. 2023 Load Forecast Supplement. https://www.pjm.com/-/media/planning/res-adeq/load-forecast/load-forecast-supplement.ashx.

[13] PJM Resource Adequacy Planning Department. 2024. PJM Load Forecast Report. https://www.pjm.com/-/media/library/reports-notices/load-forecast/2024-load-report.ashx.

[14] William Driscoll. 2022. California law would target 90% renewable and zero-carbon electricity by 2035. https://pv-magazine-usa.com/2022/09/06/california-law-would-target-90-renewable-and-zero-carbon-electricity-by-2035/.

[15] Daniel Nishball Dylan Patel and Jeremie Eliahou Ontiveros. 2024. AI Datacenter Energy Dilemma - Race for AI Datacenter Space. https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race.

[16] Economist. 2023. There's AI in them thar Hills. https://www.economist.com/business/2023/05/29/nvidia-is-not-the-only-firm-cashing-in-on-the-ai-gold-rush. *Economist* (May 2023).

[17] Virginia Electric and Power Company. 2023. 2023 Integrated Resource Plan. https://cdn-dominionenergy-prd-001.azureedge.net/-/media/pdfs/global/company/2023-va-integrated-resource-plan.pdf.

[18] Maureen Farrell and Rob Copeland. 2024. Saudi Arabia Plans $40 Billion Push Into Artificial Intelligence. https://www.nytimes.com/2024/03/19/business/saudi-arabia-investment-artificial-intelligence.html

[19] Brian Gerke, Giulia Gallo, Sarah Smith, Jingjing Liu, Peter Alstone, Shuba Raghavan, Peter Schwartz, Mary Ann Piette, Rongxin Yin, and Sofia Stensson. 2020. The California demand response potential study, phase 3: final report on the shift resource through 2030. (2020).

[20] Greenpeace. 2019. Clicking Clean Virginia: The Dirty Energy Powering Data Center Alley. https://www.greenpeace.org/usa/reports/click-clean-virginia/.

[21] EirGrid Group. 2022. Ireland Capacity Outlook 2022–2031. https://www.eirgridgroup.com/site-files/library/EirGrid/EirGrid_SONI_Ireland_Capacity_Outlook_2022-2031.pdf.

[22] EirGrid Group. 2023. Grid Alerts Explained. https://www.eirgridgroup.com/the-grid/gridalerts/

[23] EirGrid Group. 2024. Generation Capacity Statement 2023–2032. https://cms.eirgrid.ie/sites/default/files/publications/19035-EirGrid-Generation-Capacity-Statement-Combined-2023-V5-Jan-2024.pdf.

[24] Jimmy Horn, Yutong Wu, Ali Khodabakhsh, Evdokia Nikolova, and Emmanouil Pountourakis. 2020. The Long-Term Cost of Energy Generation. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (Virtual Event, Australia) *(e-Energy '20)*. Association for Computing Machinery, New York, NY, USA, 74–85. https://doi.org/10.1145/3396851.3397685

[25] Mordor Intelligence. 2022. Northern California Data Center Market Size. https://www.mordorintelligence.com/industry-reports/northern-california-data-center-market/market-size

[26] California ISO. 2023. Special Report on Battery Storage. http://www.caiso.com/Documents/2022-Special-Report-on-Battery-Storage-Jul-7-2023.pdf

[27] California ISO. 2024. 2024-2025 Transmission Planning Process Unified Planning Assumptions and Study Plan. http://www.caiso.com/InitiativeDocuments/Draft-Study-Plan-2024-2025-Transmission-Planning-Process.pdf

[28] Peter Judge. 2022. Dominion Energy admits it can't meet data center power demands in Virginia. https://www.datacenterdynamics.com/en/news/dominion-energy-admits-it-cant-meet-data-center-power-demands-in-virginia/

[29] Peter Judge. 2022. EirGrid pulls plug on 30 Irish data center projects. https://www.datacenterdynamics.com/en/news/eirgrid-pulls-plug-on-30-irish-data-center-projects/

[30] Saijel Kishan and Josh Saul. 2024. AI Needs So Much Power That Old Coal Plants Are Sticking Around. *Bloomberg* (Jan 2024). https://www.bloomberg.com/news/articles/2024-01-25/ai-needs-so-much-power-that-old-coal-plants-are-sticking-around?embedded-checkout=true

[31] Jacob Knutson. 2023. The U.S. power grid isn't ready for climate change. https://www.axios.com/2023/06/20/us-power-grid-climate-change-extreme-weather-electricity

[32] Clifford Krauss, Manny Fernandez, Ivan Penn, and Rick Rojas. 2021. How Texas' Drive for Energy Independence Set It Up for Disaster. https://www.nytimes.com/2021/02/21/us/texas-electricity-ercot-blackouts.html

[33] Seher Dareen Laila Kearney and Deep Kaushik Vakil. 2024. US electric utilities brace for surge in power demand from data centers. https://www.reuters.com/business/energy/us-electric-utilities-brace-surge-power-demand-data-centers-2024-04-10/.

[34] Tan N Le, Zhenhua Liu, Yuan Chen, and Cullen Bash. 2016. Joint capacity planning and operational management for sustainable data centers and demand response. In *Proceedings of the Seventh International Conference on Future Energy Systems*. 1–12.

[35] Kif Leswing. 2023. Nvidia's A100 is the $10,000 chip powering the race for A.I. https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html

[36] Liuzixuan Lin and Andrew A Chien. 2023. Adapting Datacenter Capacity for Greener Datacenters and Grid. In *Proceedings of the 14th ACM International Conference on Future Energy Systems*. 200–213.

[37] Minghong Lin, Adam Wierman, Lachlan LH Andrew, and Eno Thereska. 2012. Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Transactions on Networking* 21, 5 (2012), 1378–1391.

[38] Hassan Mujtaba. 2023. GPU Market Rebounds In Q2 2023: AMD, NVIDIA, and Intel See Increased Shipments, Discrete GPU Up By 12.4%. https://wccftech.com/gpu-market-rebounds-q2-2023-amd-nvidia-intel-increased-shipments-discrete-gpus-up/

[39] NVIDIA. 2023. NVIDIA A100. https://www.nvidia.com/en-us/data-center/a100/

[40] NVIDIA. 2023. Nvidia announces financial results for second quarter fiscal 2024 | NVIDIA Newsroom. https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-second-quarter-fiscal-2024

[41] Electric Reliability Council of Texas. [n. d.]. Large Flexible Load Task Force (LFLTF). https://www.ercot.com/committees/tac/lfltf.

[42] Electric Reliability Council of Texas. 2023. 2023 ERCOT System Planning Long-Term Hourly Peak Demand and Energy Forecast. https://www.ercot.com/files/docs/2023/01/18/2023-LTLF-Report.pdf

[43] Electric Reliability Council of Texas. 2023. Report on the Capacity, Demand and Reserves (CDR) in the ERCOT Region, 2024-2033. https://www.ercot.com/files/docs/2023/05/05/CapacityDemandandReservesReport_May2023_Revised2.pdf

[44] Brad Plumer and Nadja Popovich. 2024. A New Surge in Power Use Is Threatening U.S. Climate Goals. https://www.nytimes.com/interactive/2024/03/13/climate/electric-power-climate-change.html

[45] Southwest Power Pool. 2021. 2023 SPP Resource Adequacy Report. https://www.spp.org/documents/69529/2023%20spp%20june%20resource%20adequacy%20report.pdf

[46] Ana Radovanović, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, et al. 2022. Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems* 38, 2 (2022), 1270–1280.

[47] Reuters. 2024. Meta expects first shipments of new Nvidia chips later this year. https://www.reuters.com/technology/meta-does-not-expect-new-nvidia-chips-arrive-until-least-next-year-2024-03-19/.

[48] Goldman Sachs. 2023. AI investment forecast to approach $200 billion globally by 2025. *Goldman Sachs* (Aug 2023). https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html

[49] Patrick Sisson. 2024. A.I. Frenzy Complicates Efforts to Keep Power-Hungry Data Sites Green. https://www.nytimes.com/2024/02/29/business/artificial-intelligence-data-centers-green-power.html

[50] Staff. 2024. Data Centres Improved Greatly in Energy Efficiency as they Grew Massively Larger: Can this continue with AI? https://www.economist.com/technology-quarterly/2024/01/29/data-centres-improved-greatly-in-energy-efficiency-as-they-grew-massively-larger".

[51] Juveria Tabassum. 2024. Microsoft to expand its AI infrastructure in Spain with $2.1 billion investment. https://www.reuters.com/technology/microsoft-expand-its-ai-infrastructure-spain-with-21-billion-investment-2024-02-19/

[52] Jonathan Vanian and Kif Leswing. 2023. ChatGPT and generative AI are booming, but the costs can be extraordinary. https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

[53] Robert Walton. 2023. Conventional generation outages set a record in 2022: NERC. https://www.utilitydive.com/news/conventional-generation-outages-set-a-record-in-2022-nerc/653755/

[54] Jiali Xing, Bilge Acun, Aditya Sundarrajan, David Brooks, Manoj Chakkaravarthy, Nikky Avila, Carole-Jean Wu, and Benjamin C Lee. 2023. Carbon Responder: Coordinating Demand Response for the Datacenter Fleet. *arXiv preprint arXiv:2311.08589* (2023).

[55] Muyu Xu and David Stanway. 2022. Explainer: The power crunch in China's Sichuan and why it matters. https://www.reuters.com/world/china/power-crunch-chinas-sichuan-why-it-matters-2022-08-26/

[56] Fan Yang and Andrew A. Chien. 2017. Large-scale and Extreme-Scale Computing with Stranded Green Power: Opportunities and Costs. *IEEE Transactions on Parallel and Distributed Systems* 29, 5 (December 2017).

[57] Chaojie Zhang and Andrew A Chien. 2021. Scheduling Challenges for Variable Capacity Resources. In *Workshop on Job Scheduling for Parallel Processing (JSSPP).*

A GRID RESOURCE ADEQUACY ASSESSMENT DETAILS

Figure 10 to 13 depict the grid resource plan of Dominion, CAISO, ERCOT, and SPP. The diversity of resource mix and decarbonization progress can be clearly seen. As actual wind and solar generation are usually much less than their nameplate capacity (low capacity factors), the grid load limit is also much less than the total capacity in power grids with high fraction of renewable generation. For each time interval in a year, we scale 2022's renewable generation using the ratio of future capacity to 2022 capacity to model future renewable variation. We also separate non-DC load from load trace and scale it using the grid forecast to model future load variation.

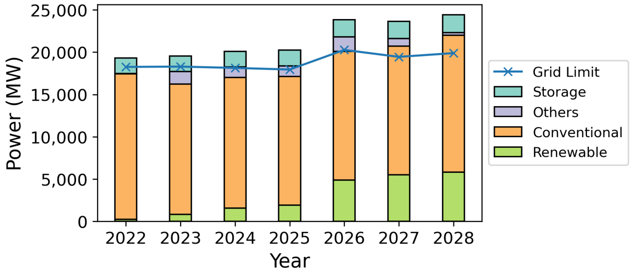


Figure 10: Dominion Reliable Load Limit and Resources.

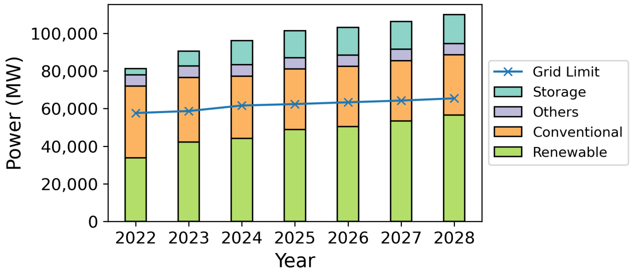


Figure 11: CAISO Reliable Load Limit and Resources.

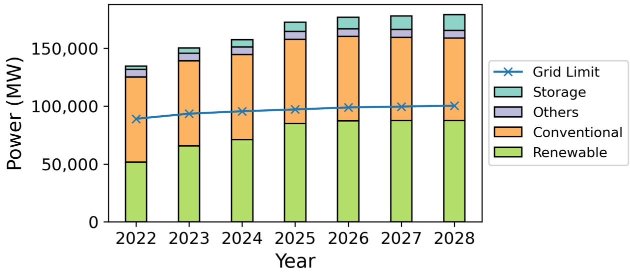


Figure 12: ERCOT Reliable Load Limit and Resources.

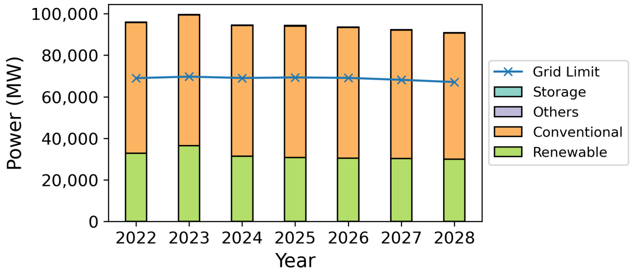


Figure 13: SPP Reliable Load Limit and Resources.

Grid resource adequacy assessment considers the uncertainty of available resources. A conventional generator is not always available because of forced outages, maintenance, etc. Energy storage only shifts power, so how much it can discharge during grid peak hours depends how it is operated. Therefore, these resources are de-rated in adequacy assessment. De-rating factors in different grids are listed in Tabel 1. We apply CAISO's 0.73 energy storage de-rating factor to other grids as we think it's good practice and mainly depends on grid regulation that others can also adopt, but note that this could make the assessment results for ERCOT and EirGrid more optimistic than the reality. For the other resources, we directly use the de-rated values in grid reports.

Table 1: De-rating Factors for Conventional Generation and Energy Storage of Different Grids

Grid	Conventional	Energy Storage
EirGrid	0.75 [21]	0.73
Dominion	0.915 [53]	0.73
CAISO	0.915 [53]	0.73 [26]
ERCOT	0.89 [43]	0.73
SPP	0.894 [45]	0.73

B DATACENTER LOAD GROWTH FORECAST

In this section, we detail our model for projecting future demand on power grids from datacenter growth. Emerging generative AI applications are expected to accelerate the cloud-driven growth of datacenters.

We project the incremental growth from AI based on NVIDIA's quarterly revenue forecast and history from its latest earnings call [40]. Under a flat annualization of quarterly revenue, and considering the price of A100 hardware ($10,000) [35], we derive the number of GPUs sold by NVIDIA. In accordance with the methodology in [6], we then forecast the corresponding datacenter capacity backed by GPUs sold by NVIDIA, assuming 300W [39] average incremental load (GPU at 50% of TDP, server balance of 50%, and datacenter PUE of 1.0). Using analyst estimates of NVIDIA's GPU market share of 68% [38], we get AI's increase in the global datacenter capacity. By normalizing this increment to a median estimate of global datacenter capacity (33,000 MW) [1], we arrive at the impact of GPU sales increase on datacenter growth—4.12% increase in CAGR.

Table 2: AI Demand-driven Datacenter Growth Estimated from NVIDIA Revenue

Quarter Ending	Oct '23	April '23
DC revenue ($B)	12	4.28
Annualized revenue($B)	48	17.12
Number of GPUs sold (M)	4.8	1.712
Resulting DC capacity (MW)	1,440	513.6
Resulting DC capacity—Global (MW)	2,117.65	755.29
Normalized to existing DC capacity	6.40%	2.28%
Incremental Global DC capacity growth	4.12%	

The incremental growth from AI is then added to the pre-AI datacenter growth. The average global datacenter growth is 7.21% CAGR, as estimated from IEA's 2015–2022 data [1]. However, certain regions see higher growth because of suitable conditions for datacenters (e.g., networking, operation cost). We use grid-projected datacenter CAGR when available. For ERCOT and SPP, grid-projected DC forecasts are not available. We consider the ERCOT region (Texas) as a future hotspot for datacenters [5] that will see growth similar to Dominion's, and use the global average DC growth for SPP. We adjust the incremental growth using:

$$CAGR_{cloud+AI,r} = CAGR_{cloud,r} + \frac{CAGR_{cloud,r}}{CAGR_{cloud,avg}} \cdot \Delta CAGR_{AI,avg}$$

that assumes AI-demand incremental capacity in a region is proportional to the region's share in global datacenter capacity growth. The base Cloud growth rates and AI-accelerated growth are listed in Table 3.

Table 3: Regional 2022 Datacenter Capacity and Estimated Growth Rate

Grid	2022 Capacity	Cloud Growth	Cloud+AI
Dominion	2767 MW [17]	15% [17]	23.56%
ERCOT	2332 MW [3]	15%	23.56%
EirGrid	600 MW [21]	10.6% [21]	16.65%
CAISO	993 MW [25]	8.5% [25]	13.35%
SPP	1810 MW [3]	7.21%	11.33%

B.1 Revisiting AI Growth

In this section, we detail the updated linear DC load growth model described in Section 5. We extend our AI load growth projections to consider NVIDIA's latest DC revenue for the quarters ending Oct '23 (actual, as opposed to the NVIDIA's forecast in section B), Jan '24 (actual) and April '24 (NVIDIA's forecast). Table 4 shows published NVIDIA DC revenue and the growth in revenue between consecutive quarters. We calculate the average quarterly growth among the post-AI boom quarters, and extrapolate NVIDIA's DC revenue for subsequent quarters, assuming linear growth in revenue amounting to $4.34B each quarter. The reminder of the projection methodology remains as described in section B.

Table 4: NVIDIA's post-AI Boom quarterly DC revenue and growth from the previous quarter

Quarter Ending	Revenue ($B)	Growth ($B)
July '23	10.32	
Oct '23	14.51	4.19
Jan '24	18.40	3.89
April '24	23.34	4.94
Average quarterly revenue growth		4.34